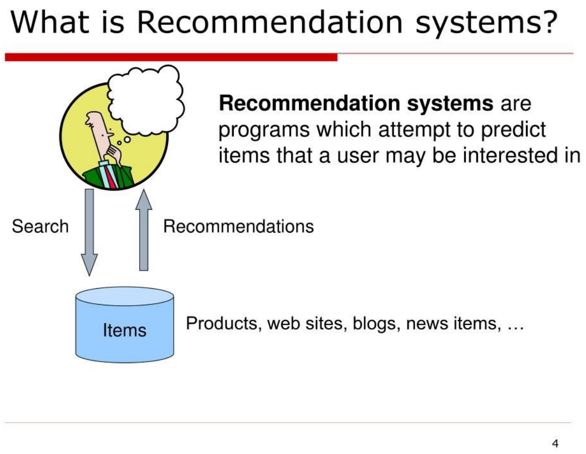
Your Next Favorite Book Awaits:
Uncover Tailored Recommendations

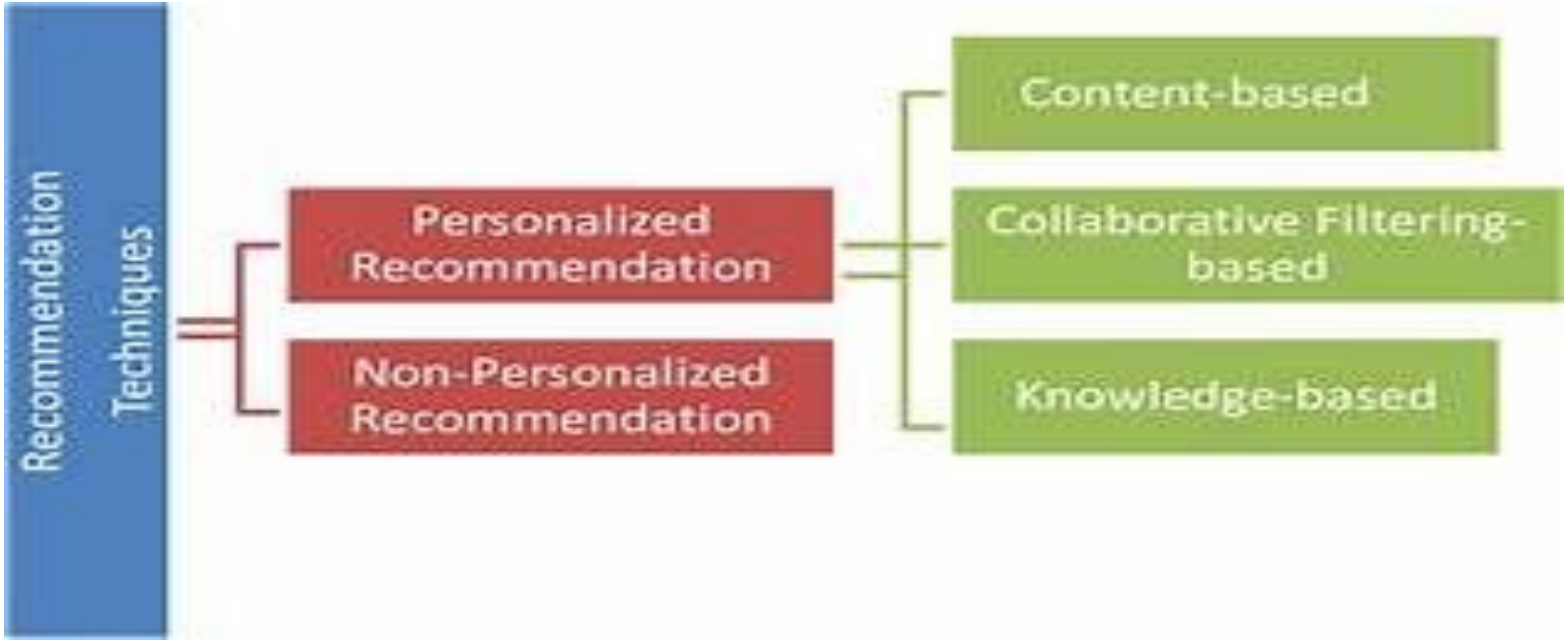
Team: Deeksha , Ankita, Fida

Why do we need a recommendation system?

Recommender systems open new opportunities of retrieving personalized information on the Internet. It also helps to alleviate the problem of information overload which is a very common phenomenon with information retrieval systems and enables users to have access to products and services which are not readily available to users on the system.



Types of Recommendation System



About the dataset

Three files:

- **Users** - Contains information about the users such as Usersid , Age , location.

- **Rating** - Contains information about books ratings such as Isbn, ratings(0-10),userid.

- **Books** - Contains information about books Such as book-title, book-author, year-of-publication, Publisher,book-image.

About the data

92 K Distinct Users
2 Lakhs Unique Books
1 Million Ratings
16 K Publishers
1 Lakhs Authors

UserId
User Age
User Location
ISBN
Rating
Book Title
Book-Author
Book-Publisher
Year of Publication

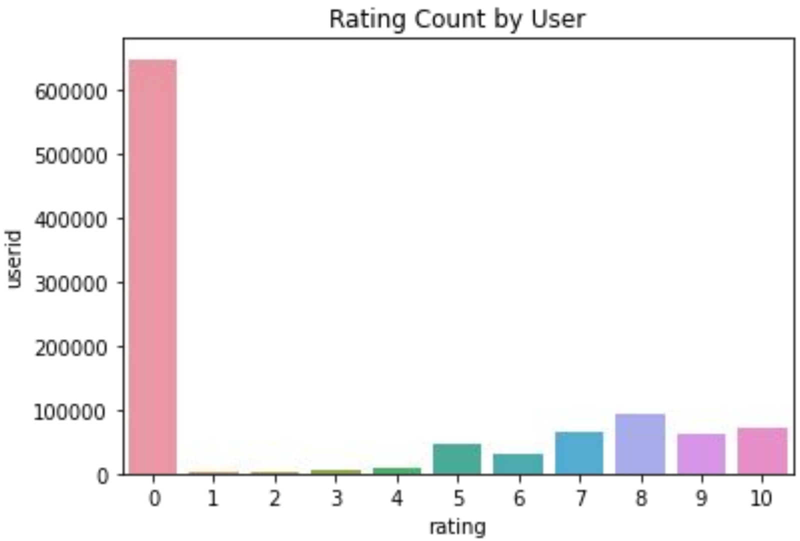
Each row is the rating given by a particular user for a specific book

Preparing Data(Data Cleanup)

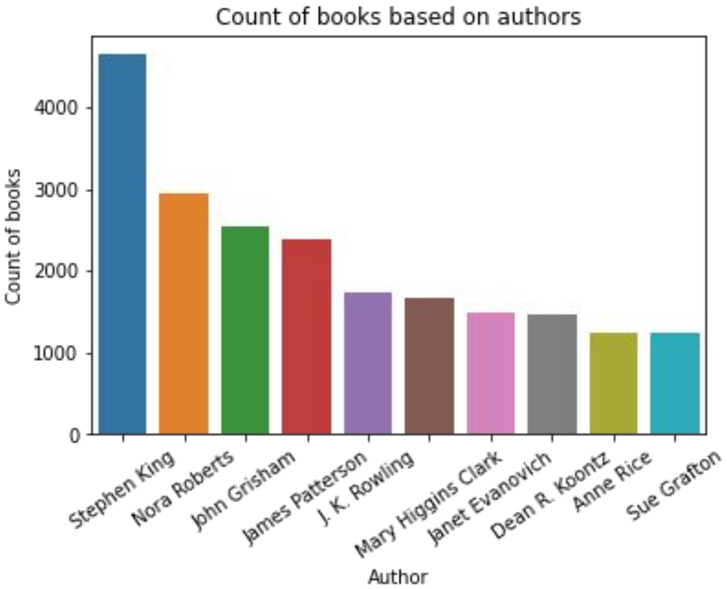
- Data was provided in 3 separate files and had to merge based on isbn and userid.
- Firstly i have checked data dimensionality (shape,size,values).
- Many null values in few columns, age column has about 26% nulls.
- Location column has country , state ,city (has few errors).
- Each file included lot of duplicates so, removed all duplicates from the data.
- Rename all the columns for the ease of understanding.

Exploratory Data Analysis

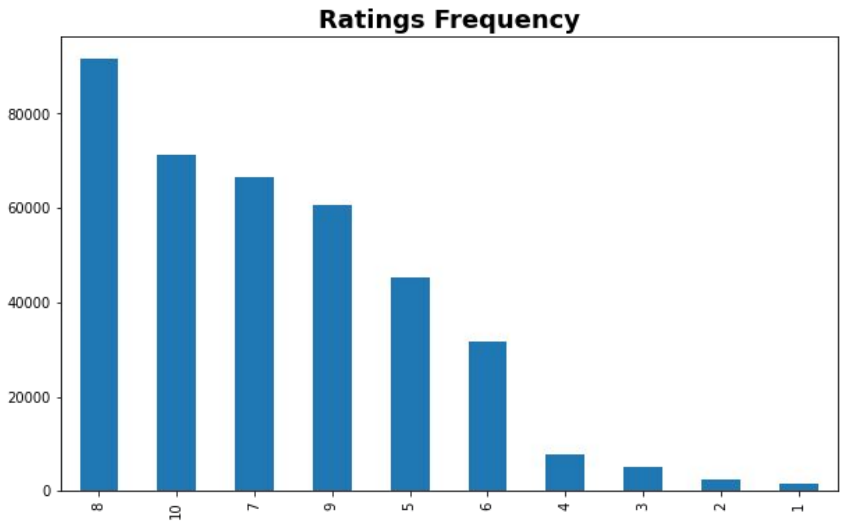
Number of Ratings By Users

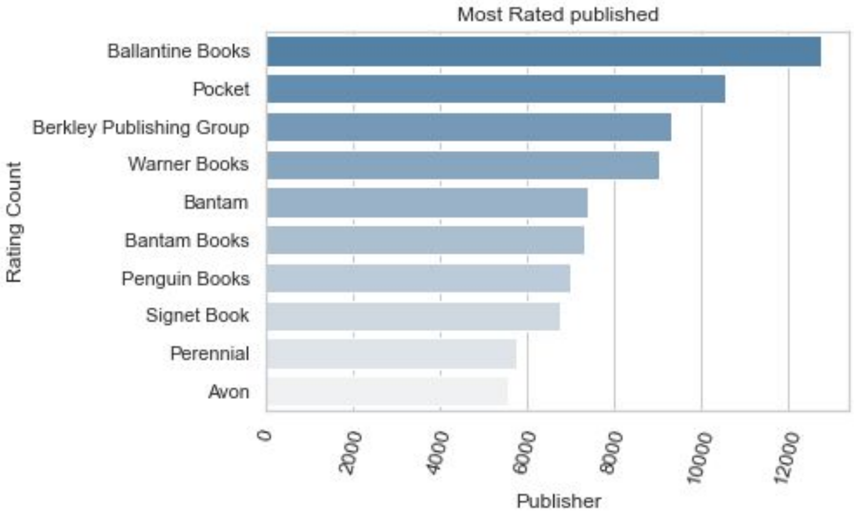


Top 10 Authors based on Ratings

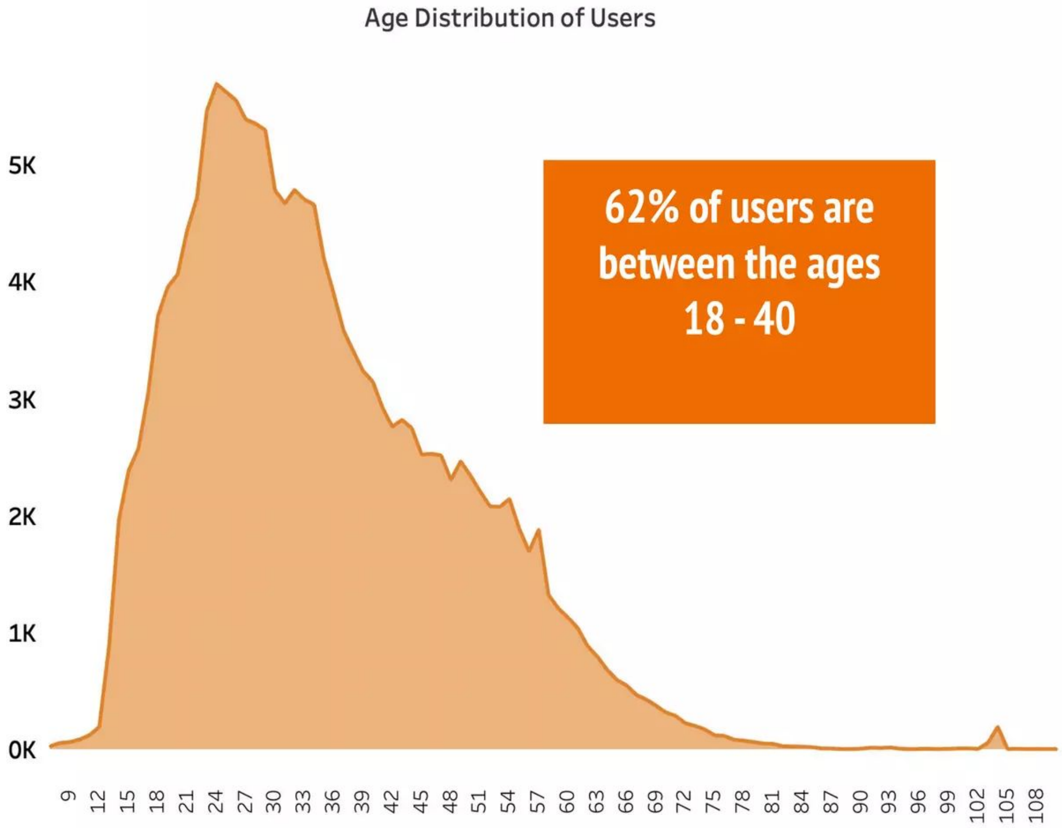


Exploratory Data Analysis



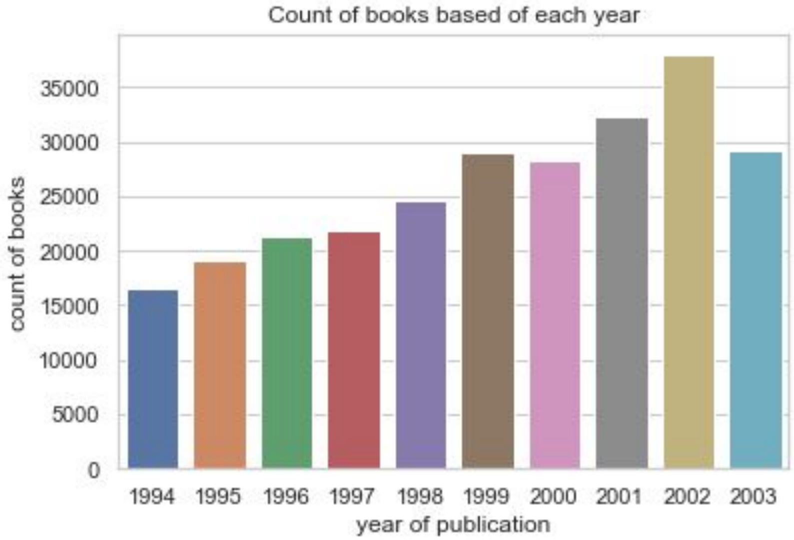


Exploratory Data Analysis

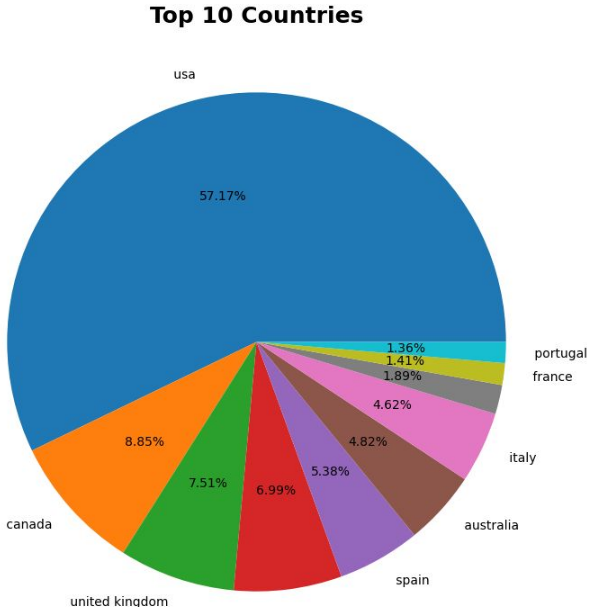


62% of users are between the ages 18 - 40

Year of Publication



Top 10 Countries

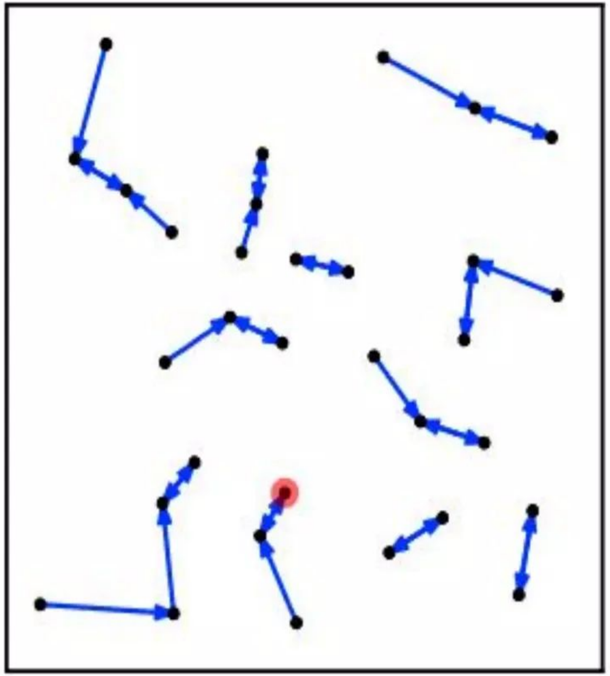


k nearest neighbors

The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to.

The principle behind nearest neighbor methods is to find a predefined number of samples closest in distance to the new point, and predict the label from these.

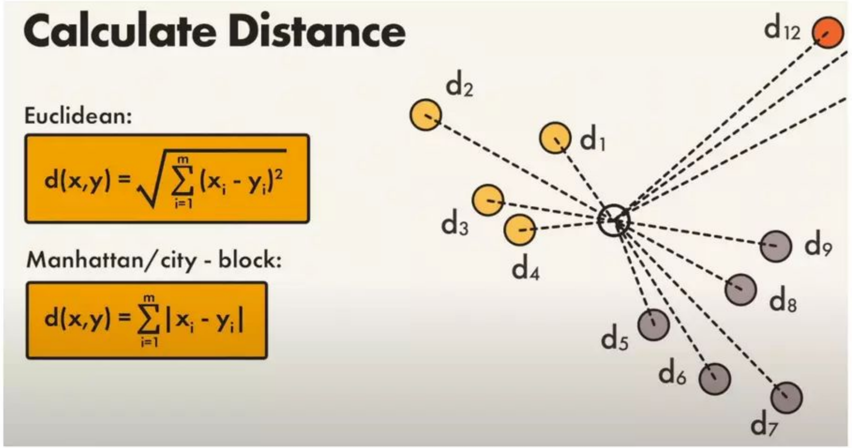
The number of samples (k) can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning).



Working Principle of KNN

- Choose the K value
- Calculate the distance between all the training points and new data points.
- Sort the computed distance in ascending order between training points and new data points.
- Choose the first K distances from the sorted list.
- Take the mode/mean of the classes associated with the distances.

For classification, compute **mode** else for regression problem compute **mean** with the distances.



Memory based Collaborative Filtering

- Filtered out users which have read at least 200 books and books which have at least 50 rating.
- Created user-item matrix.
- Now we have 706 books and 810 users.
- Calculated similarity scores for each pair of users.
- Created a function on retrieve top 5 books

```
def book(book_name):
    index=np.where(pt.index==book_name)[0][0]
    similar_items = sorted(list(enumerate(similarity_score[index])),key=lambda x:x[1],reverse=True)[1:6]

    data=[]
    for i in similar_items:
        item =[]
        temp_df= df_books[df_books['title']==pt.index[i[0]]]
        item.extend(list(temp_df.drop_duplicates('title')['title'].values))
        item.extend(list(temp_df.drop_duplicates('title')['author'].values))
        item.extend(list(temp_df.drop_duplicates('title')['Image'].values))

        data.append(item)
    return data
```

POPULARITY BASED RECOMMENDATION SYSTEM

Top 50 books with highest averge rating and people who have rated atleast 200 books

In [365]:
```
num_rating= df_books.groupby('title').count()['rating'].reset_index()
```

In [366]:
```
num_rating
```

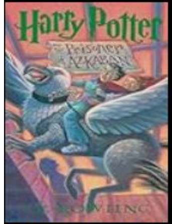
Out[366]:

	title	rating
0	A Light in the Storm: The Civil War Diary of ...	4
1	Always Have Popsicles	1
2	Apple Magic (The Collector's series)	1
3	Ask Lily (Young Women of Faith: Lily Series, ...	1
4	Beyond IBM: Leadership Marketing and Finance ...	1
...	...	...
241066	Ã?Â?lpiraten.	2
241067	Ã?Â?rger mit Produkt X. Roman.	4
241068	Ã?Â?sterlich leben.	1
241069	Ã?Â?stlich der Berge.	3
241070	Ã?Â?thique en toc	2

241071 rows × 2 columns

Top 50 Books

Top 50 Books



Harry Potter and the Prisoner of Azkaban (Book 3)

J. K. Rowling

Votes -428

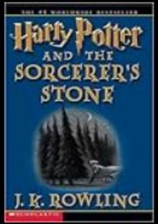
Rating - 5.852803738317757

Harry Potter and the Goblet of Fire (Book 4)

J. K. Rowling

Votes -387

Rating - 5.8242894056847545



Harry Potter and the Sorcerer's Stone (Book 1)

J. K. Rowling

Votes -278

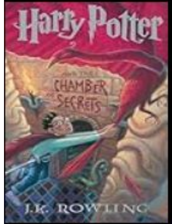
Rating - 5.737410071942446

Harry Potter and the Order of the Phoenix (Book 5)

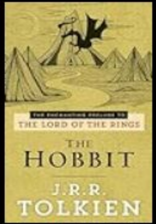
J. K. Rowling

Votes -347

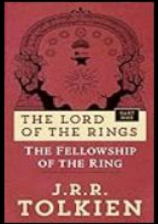
Rating - 5.501440922190202



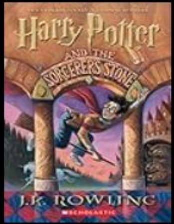
Harry Potter and the Chamber of Secrets (Book 2)



The Hobbit : The Enchanting Prelude to The Lord of the Rings



The Fellowship of the Ring (The Lord of the Rings, Part 1)



Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback)) harry

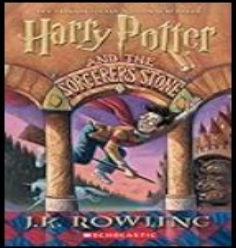
Submit

Harry Potter and the Prisoner of Azkaban (Book 3)

J. K. Rowling

Harry Potter and the Goblet of Fire (Book 4)

J. K. Rowling



Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))

J. K. Rowling

Harry Potter and the Order of the Phoenix (Book 5)

J. K. Rowling

Business Value

How Recommendation System Can Drive Business Value

- Drive customers engagement through new recommendation engine.
- Recommendation system can help make design decisions by surfacing the most rated books.

Future Improvements

- The cold-start problem: Collaborative filtering systems are based on the action of available data from similar users. If you are building a brand new recommendation system, you would have no user data to start with. You can use content-based filtering first and then move on to the collaborative filtering approach.
- Input data may not always be accurate because hall ratings are self reports. User behavior is more important than ratings.
- A strong recommendation engine will be able to identify changes (or signs of an impending changes) in customers' preferences and behavior, and constantly auto-train themselves in real time in order to serve relevant recommendations.

Thankyou